Building a 3-D Appearance Model of the Human Face

Karl Skoglund, Rasmus Larsen and Brian Lading

Informatics and Mathematical Modelling, Technical University of Denmark

September 8, 2003

Abstract

This paper describes a method for building an appearance model from three-dimensional data of human faces. The data consists of 3-D vertices, polygons and a texture map. The method uses a set of nine manually placed landmarks to automatically form a dense correspondence of thousands of points. This makes sure the model is able to capture the subtle details of a face. The model can be used for face segmentation and fully automated face registration.

1 Introduction

The human face is one of the most popular and best understood objects in research areas such as statistical shape analysis, object recognition, feature extraction, synthesis and tracking. The reason for this is the importance of the face in our daily life. The appearance of faces helps us determine properties such as gender, age, race and more intricate properties such as mood and health status. Because the face appearance is so important, the human has developed into an expert in face interpretation and recognition. This puts very high demands on an area such as face synthesis.

This article presents the design of a three-dimensional appearance model for face synthesis. An appearance model is a popular type of deformable model built from a database of examples. It is a suitable choice for addressing the problems stated above, as it generates suitable linear combinations of the examples in the training set.

Three-dimensional models are still fairly uncommon because of the large amount of overhead and computational difficulty involved in working with a three-dimensional data set. However, as computer power increases and equipment for acquiring three-dimensional data becomes reasonably priced, this type of model will be more widely used.

2 Background

In 1995 Cootes *et al* introduced the concept of shape models [5] where shapes are defined by a set of landmarks. From a database of shapes a statistical point distribution model (PDM) can be built describing the main modes of shape variation. In 1998 the shape model was extended to include texture, the intensity values contained by the shape boundaries. Such models are called *appearance models* [4]. The majority of work on appearance models have been done using two-dimensional data. The reason for this is the ease of gathering and annotating data, and the low computational power demands.

Mitchell *et al* describe the building of a three-dimensional appearance model from volumetric cardiac magnetic resonance (MR) images [11]. This paper also describes the implementation of an *active* appearance model for image segmentation and recognition. The volumetric data produced by the MR camera differs from the surface data of a human face, and the methods used cannot be used here without modifications.

Blanz and Vetter show how a three-dimensional morphable model of human faces can be built [2]. Although the model is similar to the appearance model, separate models for shape and texture are used. The dense point-to-point correspondence between shapes are formed using an optical flow algorithm. To fit the model to two-dimensional images a gradient descent optimization function is used.

Hutton *et al* build a dense correspondence model of the human face using a semi-automatic algorithm [7]. Each face is manually annotated with a sparse set of landmarks which are used to form the dense correspondence. This is the method used in this article, with a minor modification suggested by Paulsen [9].

3 Method

The data was acquired using a Minolta Vivid 900 laser scanner provided by the 3D-Laboratory at the School of Dentistry, University of Copenhagen. 15 faces were scanned, each consisting of roughly 30000 3-D points and polygons and an 800×400, 24 bit color image representing texture. Ages ranged from 20 to 40, most of the people were male and of Scandinavian origin. The age, gender and race distribution is therefore limited. One scan takes approximately 5 seconds, and three scans from different angles are necessary to get a decent representation of the face area. The scanner is not able to register hair, so a full head representation is not possible to acquire. Eyes are also hard to register since the laser beam gets too dispersed to record. Therefore, all scans are preformed with the eyes closed. The scanning process takes approximately 15 minutes per face including some post-processing.

The resulting shape and texture data is partially of poor quality. The shape is well represented but has a rough surface. This can be relaxed using a smoothing algorithm. The texture is projected onto a cylinder with resulting artifacts. The mapping from the 3-D points to the texture map (called *texture coordinates*) is poor, with many points lacking texture coordinates. This is dealt with using linear interpolation.

3.1 Annotating Shapes

When constructing a 2-D appearance model each example in the data set is annotated with corresponding landmarks. This is often done manually, although algorithms for automatic annotation exist. [1]. Around 60 landmarks is a suitable amount for a 2-D facial image. In three dimensions, thousands of landmarks are required to capture the complex surface of a human face. Obviously, it is not feasible to annotate these by hand, some sort of automated process is necessary. A semi-automatic algorithm is used here, which constructs a dense distribution of corresponding points from a sparse set of manually placed landmarks [7]. The algorithm uses one of the examples as a *template shape*. This example should be well represented and contain no statistical abnormalities. The goal of the algorithm is to change the extent and point ordering of all the other examples to that of the template. This brings all points defining the shape into correspondence. The template is first pruned so that the resulting extent of the template is present in all the other shapes. Each shape is then manually landmarked using a 3-D annotation software developed by Rasmus Paulsen which has been altered to be able to work with the type of data used here. Nine landmarks were used. These are (in this order and from the observers point of view): the chin, the left and right corner of the mouth, the tip of the nose, the left and right corner of the left eye, the nose curve minimum and the left and right corner of the right eye.

3.2 Registration

The shapes are brought into correspondence (called *registration*) as follows:

The template face is deformed [9] to roughly fit the shape to be registered using a thin-plate spline warp [3]. The two sets of landmarks define the warp transform. This makes the shapes coincide at the landmarks. The template is now treated as a set of points, and the shape to be registered is treated as a continuous surface. For each point on the template, the closest point on the target surface is found and stored. When this is done for all template points, the old target points are discarded and replaced by the new stored points. This makes sure that the template and the new shape has the same number of points and the same point ordering. An effect of this is that the new shapes are also pruned according to the template.

The template shape also defines the texture coordinates to be used in the final model. Therefore, the texture maps of all examples must be warped and pruned according to the template. The three-dimensional shape landmarks can be transformed to two-dimensional texture map landmarks by use of the texture coordinates of each landmark. These landmarks are then used to thin-plate spline warp

each texture map to fit the template. The closed-loop boundary of the shape is used to form a texture boundary. All texture maps are cropped so that only textural information inside the boundary is saved.

Since the registration makes sure each point roughly represent the same part of the face in all examples, the polygon definition of the template shape can be used for all shapes.

3.3 Procrustes Analysis

The faces now have an identical representation in both shape and texture, but before any statistical analysis can be preformed, a *Generalized Partial Procrustes Analysis* [6] must be carried out for both shapes and textures. For the shapes this means that differences in location and rotation are filtered out to leave only size and shape. Differences in size are not filtered out since the distance from the camera to the objects was held constant throughout the data acquisition. Size is an interesting attribute of a face, so it is desirable to maintain the size differences in the model. The textures are altered using a one-dimensional Procrustes analysis, making sure differences in intensity and color balance are removed.

3.4 Building the Appearance Model

A 3-D shape consisting of n_s points can be seen as a single vector in $\mathbf{R}^{3n_s}$ where the vector can be constructed as

$$\mathbf{s} = (x_0, \ldots, x_{n_s}, y_0, \ldots, y_{n_s}, z_0, \ldots, z_{n_s})$$

The set of k shapes now consists of k points in $3n_s$-dimensional space. A similar vector can be constructed for each texture map, consisting of n_t pixels. Each pixel is represented by three color values, red, green and blue, which yields the following vector representation:

$$\mathbf{t} = (r_0, \ldots, r_{n_t}, g_0, \ldots, g_{n_t}, b_0, \ldots, b_{n_t})$$

The centroids of the point clouds defined by the shape and texture vectors represent the mean shape and mean texture.

$$\bar{\mathbf{s}} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{s}_i, \quad \bar{\mathbf{t}} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{t}_i$$

The total number of dimensions of the model are $3n_s + 3n_t$. To reduce the dimensionality to something more manageable, a principal component analysis (PCA) [10] is preformed. A PCA transform rotates the coordinate system of high-dimensional data so that the axes point in directions of maximum variance. The axes are the eigenvectors of the covariance matrix of the data. For the shapes the covariance is constructed as

$$\mathbf{\Sigma_s} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T$$

The texture covariance matrix is constructed similarly. The PCA operations results in one shape model and one texture model.

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{\Phi}_s\mathbf{b}_s, \quad \mathbf{t} = \bar{\mathbf{t}} + \mathbf{\Phi}_t\mathbf{b}_t$$

The columns of the matrix $\mathbf{\Phi}$ are the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{b}$ denotes the parameters of the deformable model. To remove any correlation between the shape and texture models, a third PCA is preformed as follows. For each example in the training set the corresponding parameters are found:

$$\mathbf{b}_s = \mathbf{\Phi}_s^T(\mathbf{s} - \bar{\mathbf{s}}), \quad \mathbf{b}_t = \mathbf{\Phi}_t^T(\mathbf{t} - \bar{\mathbf{t}})$$

For each example, the shape and texture parameters are concatenated into a single vector,

$$\mathbf{b} = \begin{bmatrix} \mathbf{W}_s\mathbf{b}_s \\ \mathbf{b}_t \end{bmatrix}$$

where $\mathbf{W}_s$ is a diagonal matrix of weights accounting for the difference in magnitude between the distance units of the shapes and the intensity units of the texture. The PCA is then applied to these vectors giving a third model,

$$\mathbf{b} = \mathbf{\Phi}_c\mathbf{c}$$

Here, $\mathbf{c}$ denotes the *appearance* parameters, controlling both shape and texture.

3.5 Synthesis of Faces

By altering $\mathbf{c}$, new faces can be synthesized. $\mathbf{c} = \mathbf{0}$ results in the mean shape with the mean texture. A suitable amount of alteration is ± 3 standard deviations. From the appearance parameters $\mathbf{c}$, a new face can be synthesized by

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{\Phi}_s\mathbf{W}_s^{-1}\mathbf{\Phi}_{c,s}\mathbf{c}$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{\Phi}_t \mathbf{\Phi}_{c,t} \mathbf{c}$$

As an alternative, new faces can be synthesized using all three models by first calculating $\mathbf{b}$ from $\mathbf{c}$ and then calculating $\mathbf{s}$ and $\mathbf{t}$ from $\mathbf{b}_s$ and $\mathbf{b}_t$.

4 Results

Figure 1, 2 and 3 show the first three modes of variation. All three modes has a face size component, since the model is of size-and-shape type. Furthermore, the first mode seem to model gender, while the second and third mode model aspect ratio and amount of beard.

The six first modes each describe 10% of the total model variation. This is an effect of the low number of faces in the training set. No apparent clustering of the face vectors can occur for such a small number of examples, instead they form a roughly gaussian distribution. With more faces in the database, the PCA transform would construct a basis with clearly descending modes of variation.

The semi-automatic registration algorithm works well, but requires that the template and surface to be registered are similar. Any large variation results in uneven and incorrect registration. The problem becomes apparent when the template surface has high curvature and the novel surface curvature is low. The resulting surface will have an uneven point distribution and the high curvature parts will be cut off. For the human face, this problem occurs mainly around the nose and eyebrows. A method for regularizing correspondences found through methods such as this one is described in [8].

To be able to examine the model easily, a graphical user interface has been built, allowing the user to rotate, zoom and change the modes of variation interactively. Figure 4 shows a snapshot of the software.

5 Summary

This paper has described the building of a three-dimensional appearance model. A dense correspondence throughout all shapes was formed using nine manually defined landmarks, a thin-plate spline warp and a closest point operation. The results are satisfying, but some artifacts occur which calls for

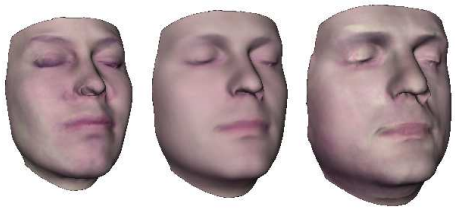

Figure 1: Mode 1 (10%). Left to right represents -3, 0 and +3 standard deviations.

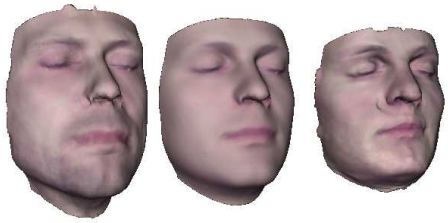

Figure 2: Mode 2 (10%)

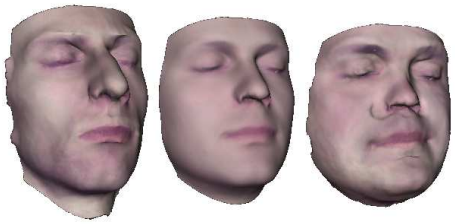

Figure 3: Mode 3 (10%)

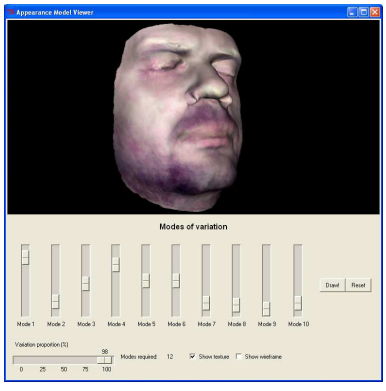

Figure 4: The appearance model viewing software.

a replacement of the point-to-surface closest point operation. To make the model more general, the database of faces needs to be extended.

6 Future Work

Currently work is being done on using the appearance model for face segmentation and fully automated face registration. The database will be extended to include more face scans. Eight new face scans exist and will be added shortly.

References

[1] Ericsson A. and strm K. An affine invariant deformable shape representation for general curves. In *Proc. 9th Int. Conf. on Computer Vision, Nice, France*, 2003 (to appear).

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In Alyn Rockwood, editor, *Siggraph 1999, Computer Graphics Proceedings*, pages 187–194, Los Angeles, 1999. Addison Wesley Longman.

[3] F. L. Bookstein. Shape and the information in medical images: A decade of the morphometric synthesis. In *Computer Vision and Image Understanding*, pages 97–118, 1997.

[4] T.F Cootes and C.J. Taylor. *Statistical Models of Appearance for Computer Vision*. University of Manchester, 2001.

[5] T.F. Cootes, C.J. Taylor, Cooper D., and Graham J. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[6] Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, 1999.

[7] Tim J. Hutton, Bernard F. Buxton, and Peter Hammond. Dense surface point distribution models of the human face, 2001.

[8] R. R. Paulsen and K. Hilger. Shape modelling using markov random field restoration of point correspondences. In *Information Processing in Medical Imaging*, IPMI, 2003.

[9] R. R. Paulsen, R. Larsen, S. Laugesen, C. Nielsen, and B. K. Ersbøll. Building and testing a statistical shape model of the human ear canal. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002, 5th Int. Conference, Tokyo, Japan,*. Springer, 2002.

[10] S.C. Sharma. *Applied Multivariate Techniques*. John Wiley & Sons, 1996.

[11] Boudewijn P.F. Lelieveldt Rob J. van der Geest Johan H.C. Reiber Milan Sonka Steven C. Mitchell, Johan G. Bosch. 3-d active appearance models: Segmentation of cardiac mr and ultrasound images. 2002.